ECRT: Evidence-Centric Cross-Modal Reasoning with Trust-Aware Gating for Multimodal Fake News Detection

Meghna Gade[1],Sriman Narayana[1], Rakesh Kumar Sanodiya[2], Lekshmi R[3],Jimson Mathew[4]
[1]IIITDM Jabalpur, Madhya Pradesh, India
[2] Indian Institute of Technology Ropar,Rupnagar, Punjab, India
[3]Muthoot Institute of Technology and Science, Kerala, India
[4]Indian Institute of Technology Patna, Bihar, India
Email: rakesh.s@iitrpr.ac.in,

***Abstract*—Fake news detection is becoming difficult due to unreliable visual evidence and inconsistent interactions between textual claims, images, and social signals. Existing methods primarily rely on feature-level fusion or static attention, lacking explicit modeling of evidence reliability. We propose ECRT (Evidence-Centric Cross-Modal Reasoning with Trust-Aware Gating), a novel framework that treats images as primary evidence signals and performs structured reasoning over claims, visual content, caption-derived semantics, and social context. ECRT introduces evidence-centric tokenization, trust-aware dynamic gating to regulate modality contributions, and a lightweight cross-modal consistency loss that improves generalization without external supervision. Experiments on the Fakeddit dataset show strong performance across multiple label granularities, achieving 93.25%, 92.95%, and 89.17% macro-F1 for 2-way, 3-way, and 6-way classification, respectively. Also on the Weibo dataset, attains 96.03% macro-F1 and 98.60% AUC. Extensive ablation studies confirm improved robustness and training stability, highlighting ECRT as an interpretable and computationally efficient alternative to many heavy multimodal approaches.**

***Index Terms*—Evidence-Centric Learning, Trust-Aware Multimodal Fusion, Fake News Detection, Vision–Language Reasoning**

I. Introduction

The increasing spread of fake news across all social media platforms will continue to degrade the quality of public dialogue and democracy [1]. Misinformation campaigns today are taking advantage of the use of multimodal content (i.e., combining textual claims and visual media) to go viral. This combination of types of inputs makes fake news detection more complex, because most indicators of deception do not occur in unimodal anomalies but rather through discrepancies in multimodal nature [2]. Consequently, effective detection requires models that not only fuse multimodal features but also conduct analytical reasoning on different evidence sources with validated levels of trust for each type. Initial attempts to detect multimodal fake news used methods to create shared representations and reduce event-specific biases using adversarial learning [3], while other recent attempts leveraged text-image similarity as deception signals but treated cross-modal consistency as static features, overlooking instance-specific modality reliability [4]. Pretrained language models like BERT further improved textual understanding [5], while vision-language pretraining enabled more robust alignment learning [6]. Additionally, image captioning techniques provided semantic bridges between modalities [7]; however, in many instances, captions have only been considered as auxiliary textual references rather than as independent evidence requiring validation.

Recent studies have explored hierarchical attention mechanisms [8] and explainable detection systems [9]. While some approaches use adaptive fusion through mixture-of-experts architectures [10] and multimodal cross-attention networks, including methods with the integration of external knowledge, affective cues, or graphs [11], [12]. Contrastive learning across modalities has shown the ability to learn robust representations [13], while retrieval-augmented methods have added high computational cost in improving contextual representations [14]. Despite these advances, existing methods have major shortcomings mainly that they make implicit assumptions on uniform reliability across modalities or learn to create attention weights in absence of an explicit model of trust, and primarily use a feature-level fusion process instead of a more logical verification of the different semantic roles of information, particularly distinguishing *claims* from *evidence*, *contradictory signals*, and *social context*. To overcome these limitations, we propose **ECRT (Evidence-Centric Cross-modal Reasoning with Trust-aware Gating)**, a unified framework for multimodal fake news detection, which introduces four key innovations:

1) **Evidence Tokenization**: Multimodal inputs are structured into four distinct semantic tokens: *Claim (C)*, *Visual Evidence (E)*, *Caption-derived Evidence (CE)*, and *Social Context (S)*; enabling role-based reasoning beyond conventional feature fusion.
2) **Trust-Aware Dynamic Gating**: A novel reasoning gate that dynamically assigns weights to each token using two complementary cross-modal consistency metrics (Claim-Evidence and Claim-Caption Similarities) and the Trust

Scores learned by the model. This adaptive weighting mechanism allows for increased emphasis on the most trustworthy evidence and suppression of misleading evidence.

3) **Caption-Mediated Alignment**: Utilize BLIP-generated captions as semantic intermediaries, explicitly calculating claim-caption alignment to reweight visual evidence.
4) **Cross-modal Consistency Loss**: A lightweight consistency constraint is utilized that measures whether the semantic content of the claims, captions, and images is in agreement. This consistency check increases the generalization ability of ECRT without requiring the utilization of any external Knowledge Bases

In fake news detection, while images are often assumed to strengthen credibility, in reality, they are often misleading or weakly related to the textual claim. Existing multimodal approaches typically assume that all modalities are equally reliable across instances. This assumption is particularly problematic when visual evidence is semantically misaligned with the claim. To analyze this issue, we computed claim-image semantic alignment using similarity scores between respective clip embeddings on two benchmark datasets. Figures 1 and 2 show the cumulative distribution of claim-image similarity for Weibo and Fakeddit datasets, respectively. A significant fraction of samples in both datasets fall below a similarity threshold of 0.4, indicating weak or misleading visual evidence. This highlights that visual content cannot be treated as uniformly trustworthy and motivates the need for explicit evidence reliability modeling. These findings directly inspire our evidence-centric design, where images are treated as primary but potentially unreliable evidence, regulated through trust-aware gating and consistency-based reasoning.

II. Related Work

Early multimodal fake news detection used adversarial training to reduce event biases [3] and cross-modal similarity as static deception signals [2]. These methods ignored instance-specific modality reliability, despite being fundamental. Stronger alignment was made possible by vision-language pre-training: CLIP-guided learning enhanced feature extraction [4], while BLIP offers semantic captioning bridges [7]. Recent fusion mechanisms include modality-interactive mixtures of experts [10] and hierarchical attention networks [8]. Although these techniques mainly carry out feature-level fusion without explicit semantic role modeling, cross-modal contrastive learning further improved discriminative power [13]. While graph-sequence models captured broad semantic relationships [12], KEN combined emotional cues and common sense knowledge [11]. Strong performance is attained by retrieval-augmented frameworks such as RADAR [14], but they come with significant computational overhead and external dependencies. Explainability initiatives like DEFEND [9] offer post-hoc justifications, but they did not incorporate trust modeling into their reasoning, which is a major drawback considering the inconsistent modality reliability of false information. Most approaches handle modalities consistently, without evaluating trust, and captions are underutilized as auxiliary text rather than as verified evidence. Also retrieval mechanisms add infeasible overhead.

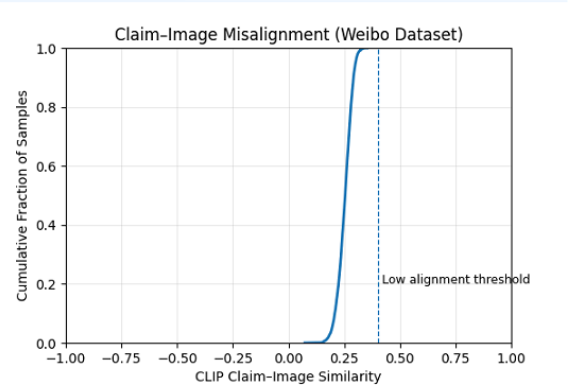

Fig. 1 Cumulative distribution of CLIP-based claim-image similarity on the Weibo dataset. A large portion of samples exhibit low semantic alignment (< 0.4), indicating unreliable visual evidence.

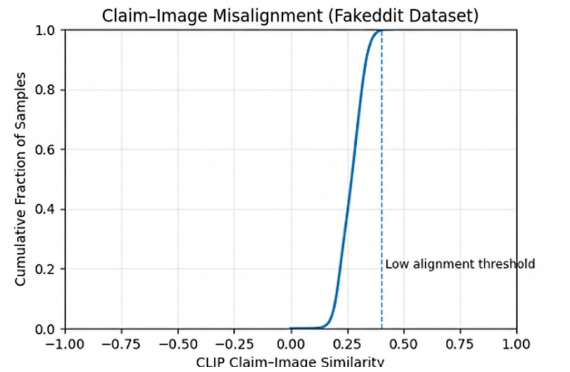

Fig. 2 Cumulative distribution of claim-image similarity on the FakeDdit dataset. Many samples fall below the alignment threshold, motivating trust-aware evidence modeling.

III. Proposed Method

We propose **Evidence-Centric Cross-Modal Reasoning with Trust-Aware Gating (ECRT)**, a multimodal fake news detection approach that explicitly separates textual claims from visual evidence and regulates their interaction through trust-aware reasoning.

A. Multimodal Feature Extraction

Every single sample given as input has a textual claim, an accompanying image, and social engagement metadata. The below details is how multimodal features are extracted:

Textual Claims : The CLIP text encoder is used to encode claims, giving normalized semantic embeddings in a shared vision-language space.

Visual Evidence : The CLIP vision encoder is used to encode images, making a direct semantic alignment with textual representations.

Caption-Derived from Evidence : First BLIP is used to generate Image captions, which convert visual content

into explicit natural-language descriptions and then encoded with CLIP text encoder to align them with claim and image embeddings in a shared semantic space. This language-level representation allows robust claim–caption consistency assessment, providing reliable visual verification beyond raw visual embeddings.

Social Metadata : To capture engagement dynamics and reduce distribution skew, these are log-transformed and standardized.

All pretrained encoders are kept frozen during training to ensure stability and efficiency. Figure 3 gives the overall workflow of ECRT framework.

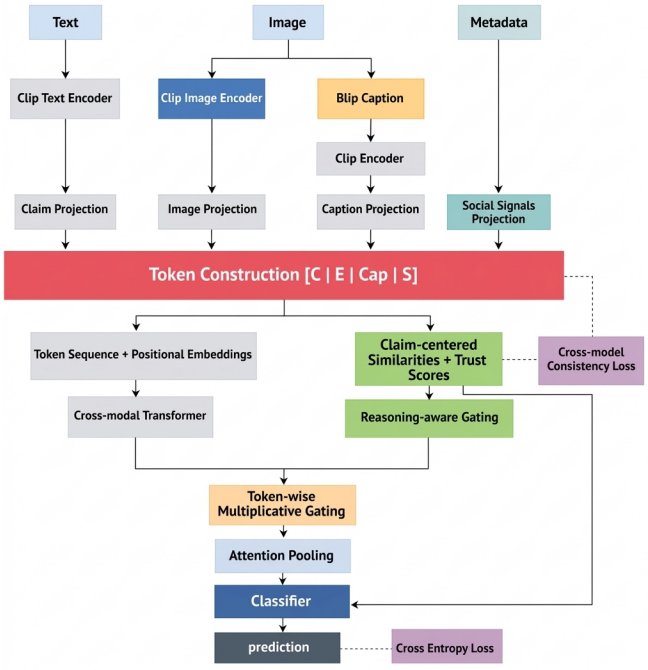

Fig. 3 Overall workflow of the proposed ECRT framework

B. Evidence Tokenization

ECRT uses four semantically different tokens to represent inputs to support structured cross-modal reasoning: a *Claim token* $\mathbf{T}_C$ capturing the textual meaning, an *Evidence token* $\mathbf{T}_E$ that represents visual content, a *Caption token* $\mathbf{T}_{Cap}$ encoding the visual evidence in a language grounded way, and a *Social token* $\mathbf{T}_S$ modeling engagement metadata. Each token is projected into the same fusion space of dimension D using modality-specific linear projections, for reliable cross-modal interactions. The resulting token sequence is added with learnable positional embeddings to maintain semantic role integrity during reasoning.

C. Caption-Aligned Evidence Refinement

Images in fake news articles frequently either fail to match the textual claim or are deliberately chosen to deceive viewers. To address this, we introduce *caption-aligned evidence refinement*, which uses visual evidence to match the semantic meaning of textual claims.

We compute a claim-caption alignment score using cosine similarity:

$$s_{\text{align}} = \cos(\mathbf{T}_C, \mathbf{T}_{Cap}) \tag{1}$$

The evidence token is refined by bounded linear modulation:

$$\mathbf{T}_E^{\text{refined}} = \mathbf{T}_E \cdot (\beta + (1-\beta)\, s_{\text{align}}) \tag{2}$$

where $\beta \in (0,1)$ is a stability coefficient controlling the minimum retained contribution of visual evidence.

This design ensures (i) stable gradient flow by preventing complete evidence suppression, (ii) soft down-weighting of semantically inconsistent images, and (iii) bounded scaling within $[\alpha, 1]$, improving training stability compared to hard gating or unnormalized attention. Captions are not treated as ground truth descriptions but as semantic validators that have language-level visual explanations.

D. Cross-Modal Transformer Reasoning

The refined token sequence is processed using a N layer Transformer encoder with multi-head self-attention, enabling contextual interaction among the 4 tokens. This allows each token to attend to all others, identifying cross-modal dependencies and contradiction patterns. The transformer outputs contextualized token representations:

$$\mathbf{H} = [\mathbf{h}_C, \mathbf{h}_E, \mathbf{h}_{Cap}, \mathbf{h}_S] \tag{3}$$

which encode cross-modal connections and are used for further fusion and classification.

E. Trust-Aware Gating Mechanism and Estimation

The process of multimodal fake news detection needs to determine how reliable different modalities are for each case. ECRT solves this problem through a trust-aware gating mechanism that measures modality trust based on initial evidence representations, before cross-modal interaction, and it uses these trust estimations together with reasoning information to control the fusion process. A lightweight trust head is used to calculate trust scores for claim and visual evidence and social context through direct application to projected token representations.

$$t_i = \sigma(\mathbf{w}_t^\top \mathbf{T}_i), \quad i \in \{C, E, S\} \tag{4}$$

Trust scores differ from cosine similarity signals: they are unary, computed per token before cross-modal interaction, capturing intra-modal feature quality rather than inter-modal alignment. They are learned implicitly via end-to-end training with cross-entropy and consistency losses, helping down-weight misleading modalities. Scores are clamped to $[0.05, 0.95]$ for stable gating across domains.

1) Reasoning-Aware Gating: To capture both semantic agreement and modality reliability, we construct a reasoning vector $\mathbf{r}$ using pre-interaction consistency signals:

$$\mathbf{r} = [\cos(\mathbf{T}_C, \mathbf{T}_E),\ \cos(\mathbf{T}_C, \mathbf{T}_{Cap}),\ \mathbf{t}] \tag{5}$$

This reasoning vector is then passed through a lightweight gating network to produce per-token modulation weights:

$$\mathbf{g} = \text{softmax}\left(\frac{\text{MLP}(\mathbf{r})}{\tau}\right) \tag{6}$$

where τ is a temperature parameter controlling gate sharpness.

2) Residual Gating with Controlled Suppression: To prevent over-suppression due to noisy trust estimates, we apply residual gating that guarantees a minimum contribution from each modality.

$$\mathbf{g}_{\text{res}} = \alpha + (1 - \alpha)\mathbf{g} \tag{7}$$

With $\alpha = 0.5$, each modality retains at least half its unmodulated contribution regardless of its estimated trust. Final gated token representations are obtained via element-wise modulation of the transformer outputs, enabling stable and adaptive multimodal fusion.

F. Attention Pooling and Classification

The attention pooling layer produces a multimodal embedding by aggregating the gated token representations via learned weights. The classifier receives as input the concatenation of this pooled representation with explicit reasoning signals (claim image similarity, claim caption similarity, and trust scores) exposing interpretable cues while preserving expressive capacity. Predictions are produced using a multilayer feedforward network with GELU activations and dropout regularization.

G. Cross-Modal Consistency Loss

To discourage reliance on spurious modality correlations, a lightweight tri-modal consistency loss is applied:

$$\mathcal{L}_{\text{cons}} = \frac{3 - (\cos(\mathbf{C}, \mathbf{E}) + \cos(\mathbf{C}, \mathbf{Cap}) + \cos(\mathbf{E}, \mathbf{Cap}))}{3} \tag{8}$$

which softly enforces semantic agreement among claim, caption, and image representations without external supervision. The overall training objective combines source and target classification loss with consistency regularization:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{cons}}\, \mathcal{L}_{\text{cons}} \tag{9}$$

where λ_{cons} controls regularization strength. The model is optimized using the AdamW optimizer with OneCycleLR scheduling and mixed-precision training for stable and efficient convergence.

IV. Experiments

We evaluate ECRT on two widely used benchmark datasets: **Fakeddit** [1] and **Weibo** [3]. Fakeddit is a large-scale multimodal dataset supporting **2-way, 3-way, and 6-way** fake news classification, using the official train/validation/test splits. Weibo consists of Chinese social media posts annotated for rumor detection. All datasets undergo strict duplicate removal across splits to avoid information leakage. For fair comparison, we evaluate ECRT against representative baselines spanning different modeling paradigms, including **text-only** models (BERT [5]), **early fusion** methods (EANN [3]), **similarity-based** approaches (SAFE [2]), **contrastive multimodal learning** (CMCL [13]), **vision–language alignment** models (CLIP-guided [4]), and **adaptive fusion** architectures (Modality MoE [10]). We also compare with recent **knowledge-enhanced** methods such as KEN [11]. Evaluation metrics include accuracy, macro F1-score, precision, recall, and AUC for binary classification tasks. All models are trained and evaluated under identical preprocessing pipelines and hardware settings. Hyperparameters are selected via ablation studies to ensure optimal and fair performance comparison. Table I and II present the dataset statistics used in our experiments.

TABLE I: Fakeddit dataset statistics under different label granularities.

Task	Class	ID	Train	Val/Test
2-Way Classification				
Real	0	36,328	5,843	5,118
Fake	1	22,920	3,701	3,239
3-Way Classification				
True	0	22,920	3,709	3,239
Partially False	1	1,532	219	233
Fully False	2	34,796	5,624	4,885
6-Way Classification				
True	0	22,920	3,709	3,239
Satire / Parody	1	3,519	560	478
Misleading	2	11,002	1,747	1,513
Imposter	3	1,223	211	168
False Connection	4	18,201	2,968	2,599
Manipulated	5	2,383	357	360

A. Implementation Details

ECRT uses CLIP-ViT-B/32 for 512D text/image embeddings, BLIP-base for caption generation (max length 30 tokens), and CLIP text encoder for caption embeddings. Social features are log-transformed and standardized. The fusion module is trained on top of frozen encoders, resulting in fast convergence and substantially lower GPU memory overhead compared to end-to-end fine-tuned multimodal models. Table III shows the metadata features used during the evaluation of ECRT with the Fakeddit and Weibo datasets. Model hyperparameters: fusion dimension $D = 128$, transformer layers $N = 2$, attention heads $h = 4$, dropout $p = 0.3$, gate temperature $\gamma = 2.0$, consistency weight $\lambda = 0.001$. Training parameters: AdamW optimizer (lr=1×10^{-3}, weight decay=0.1), batch size 128, OneCycleLR scheduler (warmup=10%), gradient clipping at norm 1.0, early stopping patience 7 epochs. Mixed precision training with GradScaler. Hardware: NVIDIA V100 32GB GPUs, PyTorch 2.0, CUDA 11.8. Reproducibility settings: seed=456, deterministic algorithms enabled, CUBLAS_WORKSPACE_CONFIG=":4096:8".

B. Results

Tables IV–VII show the performance of ECRT against multimodal baselines on Fakeddit and Weibo. On Fakeddit **2-way** classification, ECRT achieves the best overall performance with **93.25 Macro-F1** and **93.41% accuracy**, outperforming strong competitors such as KEN (93.08 F1) and Modality MoE (91.85 F1), indicating more reliable evidence utilization. For the more challenging **3-way** setting, ECRT attains the highest Macro-F1 (**92.95**) and recall (**93.05**), marginally surpassing KEN and Modality MoE, demonstrating improved

TABLE II: Class distribution of the Weibo rumour detection dataset.

Class	ID	Train	Val	Test
Non-Rumour	0	2,517	389	709
Rumour	1	2,034	283	514

TABLE III: Metadata Features Used for Dataset Evaluation

Dataset	Metadata Features Used
Fakeddit	num_comments, upvote_ratio, score
Weibo	verified, reposts, comments, likes

TABLE IV: Performance comparison on FakeDdit (2-way classification)

Model	Macro-F1	Accuracy	Precision	Recall	AUC
BERT (Text) [5]	86.88	87.53	86.85	86.90	93.79
EANN [3]	89.82	90.07	89.81	89.83	96.44
SAFE [2]	82.95	83.76	82.85	83.05	91.24
CLIP-Guided [4]	90.02	90.23	89.88	90.2	96.47
CMCL [13]	89.4	89.67	89.41	89.39	96.38
KEN [11]	93.08	93.12	92.92	93.07	98.13
Modality MoE [10]	91.85	92.02	91.71	92.01	97.57
ECRT (Ours)	**93.25**	**93.41**	**92.97**	**93.15**	**98.17**

TABLE V: Performance comparison on FakeDdit (3-way classification).

Model	Macro-F1	Accuracy	Precision	Recall
BERT (Text) [5]	85.16	86.18	86.43	84.01
EANN [3]	90.03	89.45	89.84	90.22
SAFE [2]	82.45	83.40	83.42	81.56
CMCL [13]	89.57	89.03	91.45	87.91
CLIP-Guided [4]	91.12	90.54	91.18	91.08
Modality MoE [10]	92.44	92.40	92.87	92
KEN [11]	92.82	**93.06**	92.69	92.95
ECRT (Ours)	**92.95**	92.96	**92.88**	**93.05**

class balance under increased label ambiguity. In the fine-grained **6-way** task, ECRT again leads with **89.17 Macro-F1** and **92.21% accuracy**, exceeding KEN by +0.14 F1 and outperforming CLIP-Guided and CMCL by larger margins, showing its robustness under severe class imbalance. On the Weibo dataset, ECRT achieves **96.03 Macro-F1** and **96.16% accuracy**, surpassing Modality MoE in F1 while remaining competitive in AUC. The performance of ECRT remains high and consistent across both datasets even though they vary in terms of language, platform, and label granularity.

C. Ablation Analysis

We conduct a comprehensive ablation study, which are evaluated using macro-F1 and accuracy to ensure robustness against class imbalance.

1) Component-wise Ablation Analysis: Figure 4 illustrates the effect of removing individual components from ECRT. Removing the caption-aligned evidence refinement results in a measurable performance drop, indicating that language-grounded validation of visual evidence helps suppress misleading images. Removing trust-aware gating further degrades performance, showing the importance of explicitly modeling modality reliability. Removing residual gating leads to more degradation, suggesting that unconstrained gating can over-suppress useful modalities. Finally, eliminating the Transformer encoder causes the largest performance drop, confirming the necessity of cross-modal contextual reasoning. The full model achieves the best overall performance, validating that ECRT benefits not from any single component alone.

TABLE VI: Performance comparison on FakeDdit (6-way fine-grained classification).

Model	Macro-F1	Accuracy	Precision	Recall
BERT (Text) [5]	69.19	78.28	73.95	66.27
EANN [3]	79.08	88.07	81.76	76.9
SAFE [2]	63	77.61	67.46	60.39
CMCL [13]	78.36	87.01	80.47	76.56
CLIP-Guided [4]	81.79	88.54	83.99	79.93
Modality MoE [10]	88.61	91.80	89.66	87.66
KEN [11]	89.03	92.15	90.55	**87.67**
ECRT (Ours)	**89.17**	**92.21**	**91.07**	87.59

TABLE VII: Performance comparison on the Weibo dataset (binary classification).

Model	Macro-F1	Accuracy	Precision	Recall	AUC
BERT [5]	91.89	92.23	92.96	91.27	97.95
EANN [3]	82.9	82.7	84.7	81.2	90.44
KEN [11]	94.13	94.24	94.27	94.01	98.84
CMCL [13]	83.67	84.06	83.61	83.73	91.35
Modality MoE [10]	95.63	95.75	95.71	**95.56**	**98.86**
CLIP-Guided [4]	82.89	83.24	82.75	83.08	91.16
ECRT (Ours)	**96.03**	**96.16**	**96.42**	95.72	98.60

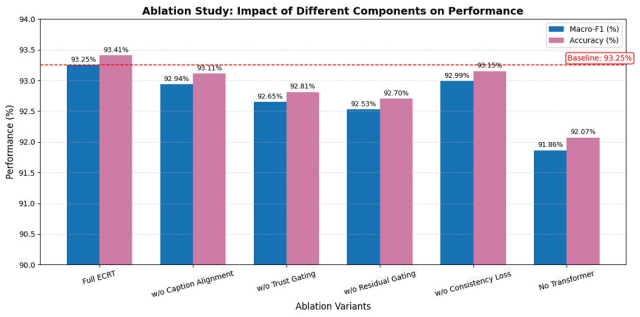

Fig. 4 Component-wise ablation study on Fakeddit (2-way)

2) Effect of Residual Gating Parameter α: To further analyze the role of residual gating, we vary the residual parameter $\alpha \in [0, 1]$, where α controls the minimum contribution of each modality after gating. Figure 5 shows that, when $\alpha = 0$, hard gating leads to aggressive modality suppression and reduced stability. Conversely, $\alpha = 1$ disables gating entirely, preventing the model from adapting to unreliable evidence. Intermediate values yield better performance. Mainly, $\alpha = 0.5$ achieves the most stable validation performance, while $\alpha = 0.7$ provides the highest peak test accuracy. This behavior confirms that residual gating acts as an implicit regularizer: it limits excessive suppression while still allowing trust-aware adaptation. Based on this, we fix $\alpha = 0.5$ in the final model to favor stable generalization across validation and test splits.

D. Effect of Cross-Modal Consistency Loss

The effect of the proposed cross-modal consistency loss is evaluated via an ablation study on the Fakeddit 3-way classification task using 3 seed values (456, 123, 789). As

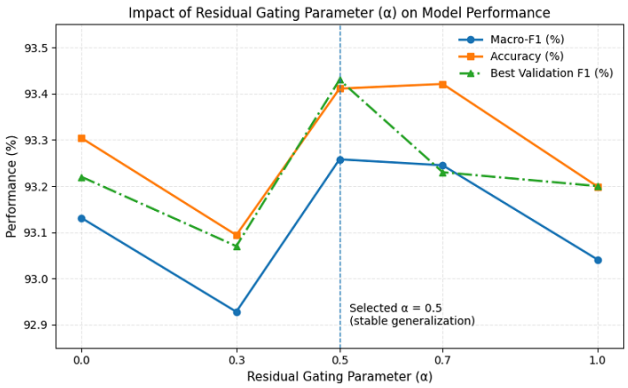

Fig. 5 Impact of residual gating parameter α on Fakeddit (2-way)

shown in Figure 6, the absolute F1 improvement is modest (+0.17%) since the primary benefit of the consistency loss is regularization, indicating improved training stability rather than raw performance gain. This behavior is consistent with its design as a lightweight constraint rather than a primary classification signal. By softly constraining cross-modal agreement, it mitigates overfitting to modality-specific features and promotes multimodal reasoning. Consequently, the consistency loss complements the trust-aware gating strategy and contributes to the robust generalization of the ECRT framework.

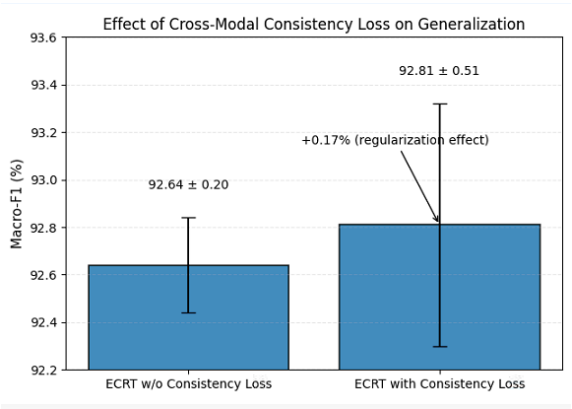

Fig. 6 Effect of cross-modal consistency loss on generalization performance on the Fakeddit 3-way classification task

V. Conclusion and Future Work

ECRT, an evidence-centric multi-modal fake news detection method, is proposed that uses structured tokenization, caption-aligned refinement, trust-aware gating and cross-modal consistency regularization to explicitly model the reliability of visual evidence. ECRT treats images as evidentiary signals instead of passive features, allowing for strong and explainable cross-modal reasoning. Evaluation on Fakeddit and Weibo datasets has shown improved results in both binary and fine-grained settings, with ablation studies indicating that these improvements are due to better evidence alignment and regularization, rather than increased model complexity. From a deployment perspective, ECRT's frozen-encoder design and lightweight fusion module make it well-suited for high-throughput social media moderation pipelines, where inference latency and computational cost are critical constraints. Future work will explore streaming adaptations and threshold calibration for production environments. ECRT can be extended to classify video-based misinformation, examine methods for adaptively aligning thresholds, and integrate external evidence retrieval into the classification process.

Acknowledgment

This work was supported by the Science and Engineering Research Board (SERB) under the Core Research Grant Project (File No. CRG/2023/001239).

References

[1] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854*, 2019.

[2] X. Zhou, J. Wu, and R. Zafarani, ": Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 2020, pp. 354–367.

[3] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 2018, pp. 849–857.

[4] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multimodal fake news detection via clip-guided learning," in *2023 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2023, pp. 2825–2830.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[8] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 153–162.

[9] L. Cui, K. Shu, S. Wang, D. Lee, and H. Liu, "defend: A system for explainable fake news detection," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 2961–2964.

[10] Y. Liu, Y. Liu, Z. Li, R. Yao, Y. Zhang, and D. Wang, "Modality interactive mixture-of-experts for fake news detection," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 5139–5150.

[11] P. Zhu, Y. Jing, L. Cheng, K. Tang, and Y. Guo, "Ken: Knowledge augmentation and emotion guidance network for multimodal fake news detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 1793–1801.

[12] J. Yin, M. Gao, K. Shu, W. Li, Y. Huang, and Z. Wang, "Graph with sequence: Broad-range semantic modeling for fake news detection," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2838–2849.

[13] X. Chen, Y. Zhang, and Q. Liu, "Cross-modal contrastive learning for multimodal fake news detection," *IEEE Transactions on Multimedia*, 2023.

[14] S.-D. Ma, Y.-H. Liu, H.-Y. Lin, P.-Y. Chen, H.-Y. Huang, S.-Y. Hsu, and Y.-N. Chen, "Radar: Retrieval-augmented detector with adversarial refinement for robust fake news detection," *arXiv preprint arXiv:2601.03981*, 2026.